Effective Prostate Cancer Detection using Enhanced Particle Swarm Optimization Algorithm with Random Forest on the Microarray Data

Original Scientific Paper

Sanjeev Prakashrao Kaulgud
Department of Computer Science, Reva University and Presidency University, Bangalore, India.
sanjeev.kaulgud@gmail.com

Siddanagouda Somanagouda Patil
Department of Applied Maths & Computer Science, University of Agricultural Sciences, Bengaluru, India
spatilsuasb@gmail.com

Vishwanath Hulipalled
Department of Computing and Information Technology, REVA University, Bengaluru, India
vishwa.gld@gmail.com

Prabhuraj Metipatil
Department of Computer Science & Engineering, REVA University, Bengaluru, India
pmetipatil@gmail.com

Abstract *– Prostate Cancer (PC) is the leading cause of mortality among males, therefore an effective system is required for identifying the sensitive bio-markers for early recognition. The objective of the research is to find the potential bio-markers for characterizing the dissimilar types of PC. In this article, the PC-related genes are acquired from the Gene Expression Omnibus (GEO) database. Then, gene selection is accomplished using enhanced Particle Swarm Optimization (PSO) to select the active genes, which are related to the PC. In the enhanced PSO algorithm, the interval-newton approach is included to keep the search space adaptive by varying the swarm diversity that helps to perform the local search significantly. The selected active genes are fed to the random forest classifier for the classification of PC (high and low-risk). As seen in the experimental investigation, the proposed model achieved an overall classification accuracy of 96.71%, which is better compared to the traditional models like naïve Bayes, support vector machine and neural network.*

Keywords*: Gene expression Omnibus, Particle Swarm Optimizer, Prostate Cancer, Random Forest*

1. INTRODUCTION

In recent decades, PC is the growing common non-cutaneous cancer type among men. According to the PC cancer agency, the death rate of PC is 2 million in 2018 [1-2]. Currently, the gene-level treatment shows great attention among the clinicians that significantly find the normal and abnormal patterns of the patients [3-4]. In Gene Expression Analysis (GEA), two major concerns in the microarray datasets are high data dimension and low samples [5]. By using the minimum number of samples, the classification of the disease may lead to incorrect decisions [6-7]. To address the above-mentioned difficulties, feature optimization is included in the GEA to find the effective subsets of features/genes [8-9]. The purpose of feature optimization is to identify the minimum number of feature subsets for attaining better classification accuracy. The conventional metaheuristics-based feature optimization algorithms like Bat optimizer, cuckoo search and artificial bee colony, include some disadvantages like it taking extra processing time, and being complex to resolve technical and scientific concerns [10]. To overcome the above-stated issues, an enhanced optimization algorithm is proposed in this manuscript to detect PC at an early stage. In this work, a new supervised system is implemented to improve the performance of PC detection using microarray data. Here, the PC-related

genes from the GEO database (GEO IDs: GSE 15484, GSE 21034, GSE 3325, and GSE 3998) were collected. The main contributions are listed below:

- After the collection of microarray data: GSE 15484, GSE 21034, GSE 3325 and GSE 3998, the gene assortment is performed utilizing an enhanced PSO algorithm on distinct GEO IDs. To enhance the searching ability of traditional PSO, many methods are included in the conventional PSO algorithm. In this article, the Interval-Newton methodology is used for keeping a reasonable search space by adjusting the swarm diversity adaptively and performing local search significantly.
- A random forest classifier is used to classify the sub-classes of PC such as low-risk (non-aggressive PC) and high-risk (aggressive PC) after selecting the optimal genes.
- The random forest classifier is the best choice for microarray data classification because it easily resolves the issue of unstructured or unbalanced data. In addition, the random forest classifier was utilized for solving both regression and classification issues, because it automatically handles the missing values in microarray data.
- The proposed model performance is related with the prior works in terms of error rate, accuracy, False Positive Rate (FPR), specificity, sensitivity and False Negative Rate (FNR).

The research paper is prepared as follows: A few research papers in PC detection using microarray data are surveyed in Section 2. The description of the proposed method is stated in Section 3. Section 4 represents the comparative and quantitative study of the work. The conclusion of this research paper is given in Section 5.

2. LITERATURE SURVEY

In recent times, microarray-based data classification gained more attention among researchers, especially for disease classification. Presently, various research works are carried out for prostate disease detection utilizing microarray data. In this section, a brief valuation of a few vital contributions to the prevailing literature is mentioned. Kim [11] developed a new inner class-clustering algorithm for classifying PC into aggressive (high-risk PC) and non-aggressive (low-risk PC). In this literature paper, the developed clustering algorithm investigates the gene pairs with a higher ranked score, which were related to the biological process of PC. In the resulting segment, the presented algorithm attained effective performance related to the existing studies by means of Area under Curve (AUC). Sharbaf [12] developed an optimization-based machine-learning scheme for prostate and Leukemia disease detection using microarray data. The developed system includes a filtering approach for reducing the dimension of the microarray data that completely lessens the system complexity and search space-time. Then, a wrapper technique was developed along with an ant colony optimizer based on cellular learning automata for extracting the feature vectors from the microarray data. Finally, the features were classified by applying three classification methods such as naïve Bayes, k-nearest neighbour, and Support Vector Machine (SVM). The experimental evaluation confirmed that the presented scheme achieved superior results by means of accuracy.

Paul and Sil, [13] presented an effective gene selection algorithm (Non-dominated Sorting Genetic Algorithm (NSGA)) for determining the biologically related genes for PC detection. The developed algorithm includes two parameters (confusion factor and risk factor). Initially, determine the risk factor of every PC-related gene to avoid data misclassification. Then, the confusion factor was calculated for each gene that helps to identify the confusion of a gene in detection, due to sample closeness in the normal and cancer classes. The experimental investigation shows that the presented gene selection algorithm attained superior performance in PC detection using specificity, accuracy, and sensitivity. Elyasigomari [14] developed a new gene selection and classification algorithm in microarray data for disease detection such as prostate, colon, leukemia, and lymphoma. In this research article, a new gene selection algorithm (cuckoo optimizer with a genetic algorithm) was developed to select the most predominant genes utilizing shuffling for better classification. After gene selection, the selected genes were classified by utilizing an artificial neural network, SVM, and Multilayer Perceptron (MP). From the experimental analysis, the SVM classifier attained better performance related to other classifiers in all the databases.

Nguyen [15] implemented a new gene selection procedure (modified analytic hierarchy procedure) for choosing the pre-dominant gene subsets for cancer DNA microarray data classification. The developed gene selection procedure chooses the relevant cancer genes based on the entropy test, Wilcoxon test, two-sample t-test, single-to-noise ratio, and receiver operating characteristics curve. Then, the selected pre-dominant gene subsets were classified using k-nearest neighbour, probabilistic neural network, Linear Discriminant Analysis (LDA), SVM, and MP. Gumaei, [16] integrated random committee ensemble learning and Correlation Feature Selection Algorithm (CFSA) for detecting PC. The experiments were performed on the public benchmark PC dataset by utilizing a tenfold cross-validation approach to analyze the developed method's performance. Alshareef, [17] developed a Chaotic Invasive Weed Optimization (CIWO) method for selecting the optimum subset of features. Further, the Deep Neural Network (DNN) model was used for detecting the existence of PC. In the present manuscript, a novel gene assortment algorithm (enhanced PSO algorithm) is proposed to improve PC recognition. Hence, the limitations and advantages of the existing papers are given in table 1.

Table 1. Limitations and the advantages of the existing papers

Author	Advantage	Limitation
Kim [11]	Developed an inner class-clustering algorithm that identifies the new unknown gene pairs, which effectively helps in differentiating the lower and higher risk cancers	The developed algorithm was only appropriate for binary classification that was considered as one of the major concerns
Sharbaf [12]	Implemented an optimization-based machine-learning scheme for prostate and Leukemia disease detection with limited computational complexity and search space time	The developed scheme was more applicable for structured data, but it showed limited results in unstructured data
Paul and Sil [13]	Developed an effective gene selection algorithm: NSGA that finds the biologically related genes for PC detection	In a few circumstances, the developed algorithm leads to class imbalance concerns
Elyasigomari [14]	Integrated cuckoo optimizer with a genetic algorithm for early disease detection such as prostate, colon, leukemia, and lymphoma	The developed gene selection algorithm needs more iterations for achieving better results which was considered as one of the key concerns in this research study
Nguyen [15]	The extensive experimental investigation shows that the LDA classifier attained good performance in all the datasets (prostate, colon, lymphoma, and leukemia cancer datasets) related to other classifiers	The undertaken classifier: LDA was mainly applicable for binary classification not for multiclass classification
Gumaei [16]	Integrated random committee ensemble learning and CFSA for early detection of PC	However, the developed method was computationally expensive

3. METHODOLOGY

In developing countries, PC is common cancer among men, which almost affects (7%) of the total population [18-19]. The major diagnostic tools applied for PC recognition are computed tomography, ultrasonic sound, magnetic resonance imaging, etc. Among these, micro-array data showed more attention among the researchers, because it represents the cell state at the molecular level [20-21]. Automatic PC detection includes a few drawbacks like limited samples and high dimensional data. In this article, a new improved gene optimization algorithm and supervised classifier are proposed for addressing the above stated drawbacks. Here, a novel supervised automated scheme is suggested for PC identification. The work proposed here contains three main stages namely data attainment, gene selection, and gene classification. The working method for the work proposed is shown in Fig. 1.

Initially, the PC-related genes are collected from the GEO dataset for automatic PC recognition. The GEO dataset is a publicly available dataset that contains original submitter-supplied records such as series, platform, and sample. Some of the series associated with PC are considered in this research article such as GSE 15484, GSE 21034, GSE 3325 and GSE 3998. A brief discussion about the undertaken GEO IDs is given as follows:

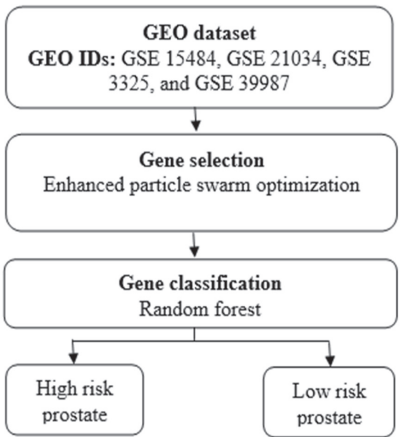

Fig. 1. Working procedure of proposed work

3.1. DATA COLLECTION

- GSE 15484 contains a total of 54675 genes. In this series, the comparison of microarray data from prostate tumors Gleason 8 (length=27), benign prostate tissues (length=13), and prostate tumors Gleason 6 (length=25).
- GSE 21034 comprises 12282 genes for 218 patients with metastatic or primary PC.
- GSE 3325 and GSE 3998 contain a total of 43419 genes and it includes thirteen individuals with metastatic and benign PC samples and six pooled samples from metastatic and benign PC tissues.
- Gene selection using enhanced PSO algorithm

Once the microarray data is collected, then the gene selection is accomplished using an enhanced PSO algorithm. Hence, PSO is a computational algorithm that mimics the behavior of a swarm rather than other evolutionary algorithms. Initially, consider a swarm size n, and then every particle i is indicated as an object with some features. Then, the particles i population are reset with position X_i and velocity V_i. Besides, estimate the objective function F_i utilizing the input value of particles position coordinates. Each particle i tracks the position coordinates, which are related to the best fitness solution achieved so far which is named Pbest. Another best value is calculated globally by the swarms, which is considered as the overall best value Gbest [22]. The particle i velocity is indicated as X_i=[X_{i1}, X_{i2},….X_{id}], and the particle i location is signified as V_i=[V_{i1}, V_{i2},….V_{id}]. In addition, every particle i has historically best position, which is specified as h_i=[h_{i1}, h_{i2},….h_{id}]. The particle's best position is identified by using the position of neighborhood particles that is specified as n_i=[n_{i1},….X_{id}]. The vectors X_0 and V_i are updated randomly by utilizing the equations (1) and (2).

$$V_{id} = wV_{id} + B_1 r_{1d}(h_{id} - X_{id}) + B_2 r_{2d}(n_{id} - X_{id}) \quad (1)$$

$$X_{id} = X_{id} + V_{id} \quad (2)$$

Where w is specified as inertia weight, B_1 and B_2 are stated as acceleration coefficients, and r_{1d} and r_{2d} are indicated as two randomly generated values within the range of [0,1] in the d dimensional space. In this article, an interval-newton method is included in the conventional PSO algorithm to keep the reasonable search space by adjusting the swarm diversity adaptively and also to perform local search effectively. The interval-newton approach is one of the effective approaches to solving a system with non-linear equations with d values. Consider F as a continuous function that has continuous partial derivatives as denoted in Eq. (3).

$$F(X^1, X^2, \dots X^d) = \begin{bmatrix} f_1(X^1, X^2, \dots X^d) \\ : \\ f_d(X^1, X^2, \dots X^d) \end{bmatrix} = 0 \qquad (3)$$

To solve the equation $F(X) = 0$, select the starting point $X_0=[X_0^1, X_0^2,\dots X_0^d]$. Then, apply an iterative formula in Eq. (3) as represented in Eq. (4).

$$X_{n+1} = X_n - J^{-1}(X_n)F(X_n),\ n = 0,1,2 \dots \qquad (4)$$

Where J is represented as the Jacobian matrix of F, and then introduce the interval–newton method $N(X)=J^{-1}(X_n)F(X_n)$ in Eq. (4), which is defined in Eq. (5) and (6).

$$X_{n+1} = X_n - N(X),\ n = 0,1,2 \dots \qquad (5)$$

$$X_{n+1} = X_n + V_{n+1} \qquad (6)$$

Where V_0 is indicated as starting velocity, X_0 is specified as starting position, V_{n+1} is represented as the present velocity of the particle i, and X_n is represented as the prior portion of the particle i. The interval-newton methodology sums the position of the particle X_n with present velocity V_{n+1}. The present velocity of particle i is identified by utilizing the acceleration constant and inertia weight, which is mathematically stated in Eq. (7).

$$V_{n+1} = wV_n + BN(X_n) \qquad (7)$$

Where, w is denoted as inertia weight, and B is indicated as acceleration constant. The gene selection after applying the enhanced PSO algorithm is denoted in Table 2.

Table 2. Gene selection after applying enhanced PSO algorithm

GEO IDs	Number of genes	Selected genes
GSE 15484	54675	300
GSE 21034	12282	300
GSE 3325 and GSE 3998	43419	300

3.3. GENE CLASSIFICATION

After gene selection, the classification of potential biomarkers is accomplished by utilizing Random Forest (RF), which solves both regression and classification issues by automatically adjusting the missing values in microarray data. RF completely lessens the issue of probability density complexity. The number of trees in an RF is stated as an individual classifier and the outcome of the classifier is chosen by all the decision trees. In this article, the tree length in the RF classification approach is fifty. The growth rule of each tree is identified to develop the RF classification approach. Then, the bootstrap samples are chosen from B for each tree in the forest, where $B(i)$ is signified as i^{th} bootstrap.

Randomly sample the micro-array data from training set S, where D is denoted as a dimension of input micro-array data, and S is stated as the number of training sets. If $d(d<D)$, select the sub-data d from the original micro-array data. From the D data, the values of d data are randomly selected and the nodes of the classifier are split by utilizing the best split on the d- dimensional data. In RF, the trees are developed until the training samples are divided without pruning. The error rate in the RF mainly depends on two dissimilar aspects:

- Power of every individual tree in the forest: The forest error rate is diminished by increasing the strength of the tree.

The connection among the trees in the forest: A higher amount of correlation leads to a higher error rate. Correspondingly, the smaller correlation leads to a smaller error rate. The pseudo-code of the proposed model is stated below:

Pseudo-code of the proposed model

Input: GSE 15484, GSE 21034, GSE 3325, and GSE 3998

Output: Classification of the potential biomarkers to predict high and low-risk prostate cancer.

Precondition: Number of relevant genes selected by enhanced PSO algorithm, number of trees in the forest F, and the training set (x_1,y_1), (x_2,y_2)..., (x_n,y_n), features (genes) D

```
Function RF (B, D)
  E ← ∅, ensemble
  For i∈1,….,F do
    B(i) ← A bootstrap section of B
    ei ← Randomized tree learn (B(i),D)
    E ← EU{ei}
  End for
  Return E
End function
Function Randomized tree learn (B, D)
  For every node:
    d ← minor subset of D
    Fragmented on top feature in d
  Return the obtained tree
End function
```

4. EXPERIMENTAL INVESTIGATION

The proposed analysis is simulated by MATLAB tool (version 2018a) in this research paper. To verify the efficacy of the proposed work, the performance of the proposed work is related with some previous works on

the GEO database. The proposed work output is validated in terms of error rate, FPR, specificity, sensitivity, accuracy and FNR. The system configuration and the parameter settings are given in table 3.

Table 3. System configuration and the parameter settings

System configuration	
Random access memory	8 GB
Hard-disk	1TB
Operating system	Windows10
Processor	Intel core i5
Enhanced PSO algorithm	
Acceleration constant	02
Initial inertia weight	0.9
Final inertia weight	0.4
Population size	300
Random forest	
Number of trees	50
Maximal depth	300

4.1. PERFORMANCE METRIC

The performance metric is described as the technique of evaluating the facts of a group or separate variable's performance by means of error rate, FPR, specificity, sensitivity, accuracy and FNR. The mathematical formula for calculating the error rate, FPR, specificity, sensitivity, accuracy and FNR are represented in Eq. (8), (9), (10), (11), (12), and (13).

$$Error\ rate = \frac{FP+FN}{P+TP+FN+TN} \times 100 \quad (8)$$

$$FPR = \frac{FP}{TN+FP} \times 100 \quad (9)$$

$$Specificity = \frac{TN}{FP+TN} \times 100 \quad (10)$$

$$Sensitivity = \frac{TP}{FN+TP} \times 100 \quad (11)$$

$$Accurcay = \frac{TP+TN}{FP+TP+FN+TN} \times 1000 \quad (12)$$

$$FNR = \frac{FN}{TP+FN} \times 100 \quad (13)$$

Where, TN, FP, TP and FN correspondingly indicate True Negative, False Positive, True Positive, and False Negative.

4.2. QUANTITATIVE INVESTIGATION

In this segment, the GEO database is applied for measuring the effectiveness of the given work. In this research, PC recognition is performed on a data mining platform for categorizing the sub-classes of PC like high and low-risk cancers. Here, the performance valuation is done with data split as 70% training and 30% testing. The performance of the suggested model is given in Table 4 and it is evaluated by considering the performance metrics like accuracy, sensitivity, and specificity for 50 iterations. Additionally, the effectiveness of the suggested model is compared with three present classification methods like SVM, Neural Network (NN), and Naive Bayes (NB). From the simulation result, the average accuracy of RF is 96.71% (combines all four GEO IDs: GSE 15484, GSE 21034, GSE 3325, and GSE 3998) and the comparative classification approaches: NN, SVM, and NB attains 72.78%, 85.05%, and 84.49% of accuracy. The average sensitivity of the RF classifier is 95.3% and the existing classification approaches achieve an average sensitivity of 81.28%, 87.22% and 85.27%. Similarly, 95.58% is the average specificity of RF and the existing classifiers (NN, SVM, and NB) attain an average specificity of 62.43%, 76.92% and 82.07%. Fig. 2 shows the Graphical representation of the suggested work in terms of accuracy, sensitivity, and specificity.

Table 4. Performance evaluation of the suggested work in terms of accuracy, sensitivity, and specificity with different classifiers

GEO IDs	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
GSE 15484	NN	66.67	67.17	68.12
	SVM	72.67	80	53.33
	NB	76.67	80	73.33
	RF	93.33	93.34	93.32
GSE 21034	NN	72.92	79.17	66.67
	SVM	87.50	91.67	83.33
	NB	80.56	83.33	77.78
	RF	98.11	94.89	95.33
GSE 3325 and GSE 3998	NN	78.75	97.5	52.5
	SVM	95	90	94.11
	NB	96.25	92.5	95.12
	RF	98.7	97.82	98.1

Table 5. Performance evaluation of the proposed work by means of error rate, FPR and FNR with dissimilar classifiers

GEO IDs	Classifier	Error rate	FPR	FNR
GSE 15484	NN	33.33	32.19	31.13
	SVM	27.67	20	46.67
	NB	23.21	20	26.67
	RF	13.33	17.27	23.67
GSE 21034	NN	27.08	20.83	33.33
	SVM	12.50	8.33	16.67
	NB	19.44	16.67	22.22
	RF	1.89	11.11	16.67
GSE 3325 and GSE 3998	NN	21.25	19.56	47.5
	SVM	5	12.23	34.4
	NB	13.75	19.5	36.5
	RF	1.3	11.7	26.6

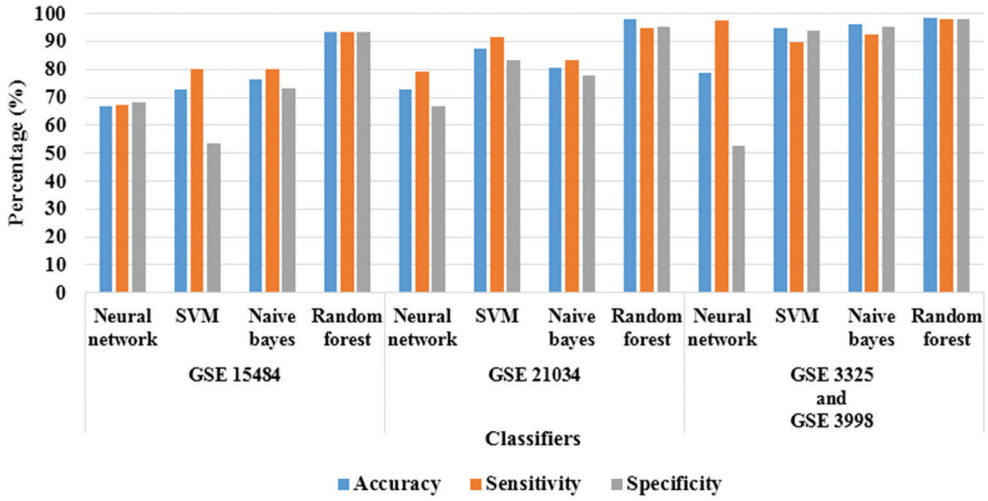

Fig. 2. Graphical representation of the suggested work in terms of accuracy, specificity, and sensitivity

In Table 5, the proposed work effectiveness is evaluated by means of error rate, FPR and FNR for 50 iterations. From the simulation consequences, the average error rate value of the RF classifier is 5.5% and the comparative classification approaches: NN, SVM, and NB classifiers achieve 27.22%, 15.05%, and 18.8% of average error rates. In addition, the FPR average value of the RF is 13.36% and existing classifiers (NN, SVM, and NB) attain 24.19%, 13.52% and 18.72% of average FPR value. At last, the FNR average value of the RF is 22.31% and the existing classifiers achieve 37.32%, 32.58% and 28.46% of the average FNR value. Tables 4 and 5 presented that the suggested work executes effectively on the GEO database (GEO IDs: GSE 15484, GSE 21034, GSE 3325, and GSE 3998) in light of error rate, FPR, specificity, sensitivity, accuracy and FNR. The graphical representation of the suggested work in terms of error rate, FPR, and FNR is shown in Fig. 3.

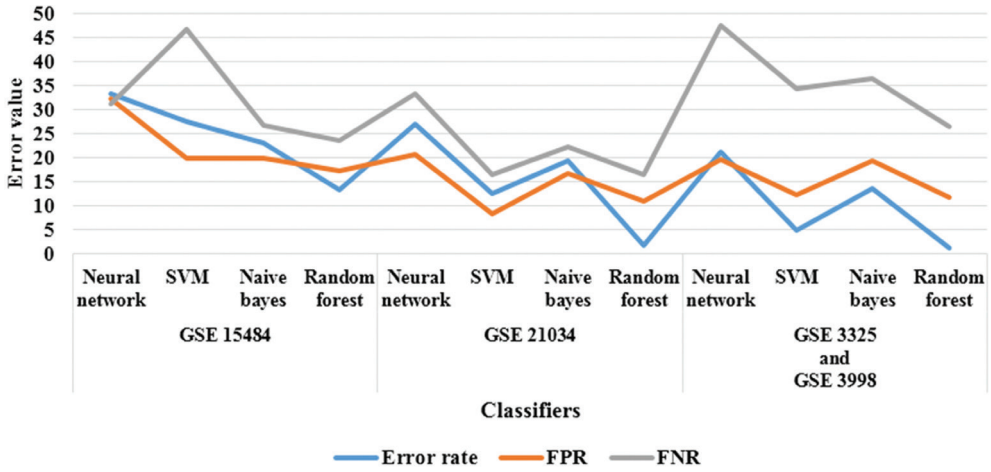

Fig. 3. Graphical representation of the suggested work in terms of error rate, FPR and FNR

Table 6 validates the performance of the suggested work with conventional PSO algorithm and enhanced PSO algorithm in light of accuracy. In the enhanced PSO algorithm, the RF averagely enhances the accuracy of PC recognition by up to 1.5-5% associated to the PSO algorithm. Here, in this work, feature optimization plays a critical role in PC recognition. Generally, the collected GEO data contains several features (genes or potential biomarkers) that might give rise to the "curse of dimensionality" problem. Therefore, reducing the dimensionality is important to optimize the genes or to select the ideal genes (associated with PC), which are appropriate for superior classification. The effectiveness of reducing dimensionality is shown in Table 6 and Fig. 4.

Table 6. Comparison of performance of the proposed work with different optimization algorithms

GEO IDs	Gene selection	Classifier	Accuracy (%)
GSE 15484	PSO	RF	90.17
	Enhanced PSO		93.33
GSE 21034	PSO		93.72
	Enhanced PSO		98.11
GSE 3325 and GSE 3998	PSO		96.29
	Enhanced PSO		98.71

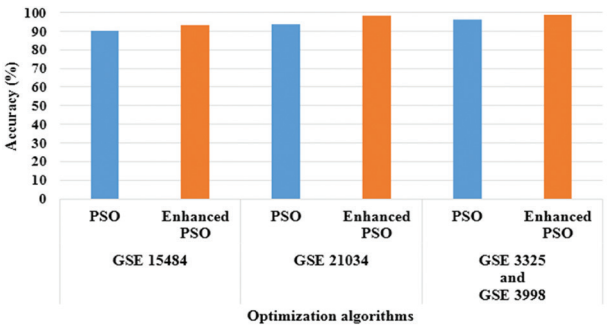

Fig. 4. Graphical illustration of the proposed work with different optimizers

4.3. COMPARATIVE ANALYSIS

In this section, the comparative evaluation of the proposed and the existing works are shown in tables 7 and 8. Kim [11] implemented a novel inner class-clustering algorithm for classifying the sub-classes of PC as aggressive or non-aggressive. In this study, the developed algorithm was tested on GEO dataset (GEO IDs: GSE 15484 and GSE 21034). From the experimental simulation, the developed algorithm attained 0.876 of AUC value on GSE 15484, and 0.989 of AUC value on GSE 21034. Compared to the published works, the proposed model achieved 0.923 of AUC value on GSE 15484 and 0.991 of AUC value on GSE 21034, which were high, related to the inner class-clustering algorithm.

Table 7. Comparative study of proposed and existing works in terms of AUC

Methods	GEO IDs	AUC
Inner class clustering algorithm [11]	GSE 15484	0.876
	GSE 21034	0.989
Enhanced PSO with RF	GSE 15484	0.923
	GSE 21034	0.991

Gumaei, [16] combined random committee ensemble learning model and CFSA for detecting PC. The developed model has averagely obtained 95.10% of recall and 95.09% of accuracy on the benchmark PC dataset. Correspondingly, Alshareef, [17] integrated CIWO and DNN for detecting the existence of PC. Hence, the developed CIWO-DNN model achieved 97.25% of recall and 97.19% of accuracy. However, the proposed enhanced PSO with RF model averagely obtained 95.35% of recall and 96.71% of accuracy in PC detection. As a future extension, a new hybrid feature optimization algorithm can be implemented to further improve PC detection.

Table 8. Comparative study of proposed and existing works by means of accuracy and recall

Methods	Accuracy (%)	Recall (%)
Correlation feature optimization with committee ensemble learning [16]	95.09	95.10
CIWO-DNN [17]	97.19	97.25
Enhanced PSO with RF	96.71	95.35

In the proposed research work, a new optimization algorithm (enhanced PSO) is developed for selecting the optimal PC-related genes. The enhanced PSO algorithm maintains better search space by adjusting swarm diversity that significantly reduces the computational complexity, which is the main concern mentioned in the literature section. Hence, the efficiency of the proposed optimization algorithm is given in tables 7 and 8.

5. CONCLUSION

This research aims to suggest new feature optimization (enhanced PSO) algorithm for choosing the most informative potential biomarkers for PC recognition. The selected potential biomarkers (PC-related genes) are classified by employing an RF classifier. The undertaken classification approach classifier effectively classifies the subclasses of PC: high-risk and low-risk PC. Related to the existing work, the developed work obtained an efficient performance in terms of accuracy. The proposed work achieved 96.71% of the overall classification accuracy in the GEO dataset from this research analysis, which is superior compared to the earlier research works. A new hybrid feature optimization algorithm is combined with a multi-class classifier to further improve the efficiency of PC identification in future work.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The paper background work, conceptualization, methodology, dataset collection, implementation, result analysis and comparison, preparing and editing draft, as well as visualization have been done by the first and second authors. The supervision, review of work, and project administration, have been done by the third and fourth authors.

6. REFERENCES:

[1] B. Lee et al. "Long noncoding RNAs as putative biomarkers for prostate cancer detection", The Journal of Molecular Diagnostics, Vol. 16, No. 6, 2014, pp. 615-626.

[2] P. Östling et al. "Systematic analysis of microRNAs targeting the androgen receptor in prostate cancer cells", Cancer Research, Vol. 71, No. 5, 2011, pp. 1956-1967.

[3] S. M. G. Espiritu et al. "The evolutionary landscape of localized prostate cancers drives clinical aggression", Cell, Vol. 173, No. 4, 2018, pp. 1003-1013.e15.

[4] A. A. Dmitriev et al. "Identification of novel epigenetic markers of prostate cancer by NotI-microar-

ray analysis", Disease Markers, Vol. 2015, 2015, p. 241301.

[5] E. K. Markert, H. Mizuno, A. Vazquez, A. J. Levine, "Molecular classification of prostate cancer using curated expression signatures", Proceedings of the National Academy of Sciences of the United States of America, Vol. 108, No. 52, 2011, pp. 21276-21281.

[6] T. Mehmood, A. Kanwal, M. M. Butt, "Naive Bayes combined with partial least squares for classification of high dimensional microarray data", Chemometrics and Intelligent Laboratory Systems, Vol. 222, 2022, p. 104492.

[7] S. Begum, R. Sarkar, D. Chakraborty, S. Sen, U. Maulik, "Application of active learning in DNA microarray data for cancerous gene identification", Expert Systems with Applications, Vol. 177, 2021, p. 114914.

[8] S. Sucharita, B. Sahu, T. Swarnkar, "An Empirical Analysis of PCA-SVM Model for Cancer Microarray Data Classification", Advances in Intelligent Computing and Communication, Lecture Notes in Networks and Systems, Vol. 202, pp. 495-504, Springer, Singapore, 2021.

[9] J. Kim et al. "Hydrogel-based hybridization chain reaction (HCR) for detection of urinary exosomal miRNAs as a diagnostic tool of prostate cancer", Biosensors and Bioelectronics, Vol. 192, 2021, p. 113504.

[10] N. S. Mohamed, S. Zainudin, Z. A. Othman, "Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data", Expert Systems with Applications, Vol. 90, 2017, pp. 224-231.

[11] H. Kim, J. Ahn, C. Park, Y. Yoon, S. Park, "ICP: A novel approach to predict prognosis of prostate cancer with inner-class clustering of gene expression data", Computers in Biology and Medicine, Vol. 43, No. 10, 2013, pp. 1363-1373.

[12] F. V. Sharbaf, S. Mosafer, M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization", Genomics, Vol. 107, No. 6, 2016, pp. 231-238.

[13] A. Paul, J. Sil, "Identification of differentially expressed genes to establish new Biomarker for cancer prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 16, No. 6, 2019, pp. 1970-1985.

[14] V. Elyasigomari, M. S. Mirjafari, H. R. Screen, M. H. Shaheed, "Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization", Applied Soft Computing, Vol. 35, 2015, pp. 43-51.

[15] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, "A novel aggregate gene selection method for microarray data classification", Pattern Recognition Letters, Vol. 60-61, 2015, pp. 16-23.

[16] A. Gumaei, R. Sammouda, M. Al-Rakhami, H. Al-Salman, A. El-Zaart, "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression", Health Informatics Journal, Vol. 27, No. 1, 2021, p. 1460458221989402.

[17] A. M. Alshareef et al. "Optimal Deep Learning Enabled Prostate Cancer Detection Using Microarray Gene Expression", Journal of Healthcare Engineering, Vol. 2022, 2022, p. 7364704.

[18] X. Wan et al. "Identification of androgen-responsive lncRNAs as diagnostic and prognostic markers for prostate cancer", Oncotarget, Vol. 7, No. 37, 2016, pp. 60503-60518.

[19] M. Lohr et al. "Identification of sample annotation errors in gene expression datasets", Archives of Toxicology, Vol. 89, No. 12, 2015, pp. 2265-2272.

[20] B. Yang et al. "Downregulation of miR-139-5p promotes prostate cancer progression through regulation of SOX5", Biomedicine & Pharmacotherapy, Vol. 109, 2019, pp. 2128-2135.

[21] D. Wen, J. Geng, W. Li, C. Guo, J. Zheng, "A computational bioinformatics analysis of gene expression identifies candidate agents for prostate cancer", Andrologia, Vol. 46, No. 6, 2014, pp. 625-632.

[22] K. L. Du, M. N. S. Swamy, "Particle Swarm Optimization", Search and Optimization by Metaheuristics, Techniques and Algorithms Inspired by Nature, Springer International Publishing, 2016, pp. 153-173.